

Pronunciation recognition of English phonemes /ə/, /æ/, /ɑ:/ and /ʌ/ using Formants and Mel Frequency Cepstral Coefficients

Keith Y. Patarroyo and Vladimir Vargas-Calderón*

The Vocal Joystick Vowel Corpus, by Washington University, was used to study monophthongs pronounced by native English speakers. The objective of this study was to quantitatively measure the extent at which speech recognition methods can distinguish between similar sounding vowels. In particular, the phonemes /ə/, /æ/, /ɑ:/ and /ʌ/ were analysed. 748 sound files from the corpus were used and subjected to Linear Predictive Coding (LPC) to compute their formants, and to Mel Frequency Cepstral Coefficients (MFCC) algorithm, to compute the cepstral coefficients. A Decision Tree Classifier was used to build a predictive model that learnt the patterns of the two first formants measured in the data set, as well as the patterns of the 13 cepstral coefficients. An accuracy of 70% was achieved using formants for the mentioned phonemes. For the MFCC analysis an accuracy of 52 % was achieved and an accuracy of 71% when /ə/ was ignored. The results obtained show that the studied algorithms are far from mimicking the ability of distinguishing subtle differences in sounds like human hearing does.

Keywords: sound processing, formants, Mel frequency cepstral coefficients, pronunciation recognition, machine learning.

1 Introduction

Throughout computing history, scientists have developed a vast amount of theories and algorithms for speech recognition that are widely known (e.g. see the work by Lee (1988)). Many of them are motivated by using some of the principles of the human ear operation (Davis & Mermelstein, 1980). In recent years, the blooming of high-speed processing computers has allowed us to start using deep machine learning to improve the efficacy of our speech recognizers (Baker et al., 2009). The techniques that are currently used for speech recognizers yield high prediction rates (Hinton et al., 2012). The most common and successful technique is the implementation of multi-layered neural networks (Zegers, 1998; Wellekens, 1998). The performance of such techniques seeded the question of implementing recognizers as objective evaluators of speech ability in humans.

This takes great relevance in the field of pronunciation teaching, not only because it provides the teachers (especially teachers who are non-native speakers of the taught language) a tool for assessing objectively the pronunciation of their students (Neri, Mich, Gerosa, & Giuliani, 2008), and of themselves, but also because if a student has access to such a tool, he or she could learn pronunciation autonomously (Hinks, 2003). Unquestionably, the main goal in pronunciation teaching is to improve the ability of a language learner to use their vocal tract to produce sounds that are recognized as native sounds by native speakers of that language. Although this remains a challenge,

*Physics Department. National University of Colombia.
Corresponding author e-mail: vvargasc@unal.edu.co

particularly in the early stages of learning a language (Mirzaei, Gowhary, Azizifar, & Esmaeili, 2015), it also provides a means by which students can improve the learning and retention of grammatical structures (Martin & Jackson, 2016). Also, since different languages have different sound systems, it is normal for language learners to struggle learning them.

For instance, the phonemes /ə/, /æ/, /ɑ:/ and /ʌ/ found in the English language are troublesome for many English learners because of their similarity. For example, Spanish and Italian speakers tend to mix these phonemes into a single one: /a/. There are cases of learners from other languages, such as Azerbaijani, in which this confusion has been studied (Ghaffarvand Mokari & Werner, 2016).

This article aims to study the ability of two of the most historically important speech recognition tools to differentiate between these phonemes, i.e. how good tools would they be in evaluating the correct pronunciation of these phonemes. These tools are the formants (computed with Linear Predictive Coding), attractive for its simplicity in vowel pronunciation recognition, and the Mel Frequency Cepstral Coefficients, attractive for its computational speed in continuous speech recognition. Both tools were tested with the Vocal Joystick Vowel Corpus, from the Washington University.

2 Theoretical Framework

2.1 Sound Formation

The frequency range of the pressure waves in the air that compose the audible sounds by humans goes from 20Hz to 20kHz (Rosen & Howell, 2011). The process by which sound is produced by humans is the following (for a complete understanding of sound production by humans see reference (O’Grady & Dobrovolsky, 1997)).

The diaphragm and intercostal contractions make the lungs generate a flow of air from the chest to the mouth. This air first passes through the larynx, wherein the vocal cords are. These are muscles that can be geometrically distributed in several ways, each of which is excited by the passing air creating modes of vibration or glottal states that result in sound. The pharynx, the oral cavity and the nasal cavity are filters and resonators of the aforementioned sounds. Also, the tongue and lips allow us to rapidly articulate and change the shape of the vocal cavity in order to filter some frequencies.

In this study, we are interested in vowels, which are voiced glottal states produced with little obstruction of the vocal tract. This means having the lips open and also the tongue without contact with the palate.

2.2 Human Hearing Detection Principles

It is also convenient to mention some of the principles that the human ear uses to detect sound signals. The human ear analyses pressure variations in the air into different frequencies. Roughly, the human ear does a Fourier transform of the pressure signal $P(t)$, where t is the time, and transmits $|P(\omega)|^2$, where ω is the frequency, to the brain (Sethna, 2006).

However, this initial model falls short for many analyses. Some of the reasons are (Lyons, n.d.): the tonal information is changing as a word or tune progresses; the difference between two closely spaced frequencies is hard to recognize for humans, especially at high frequencies; and humans hear loudness on a non-linear scale. All of these factors propose a significant challenge to mimic the human hearing system.

Davis and Mermelstein (1980) developed a method (Mel Frequency Cepstral Coefficients) to mimic this process, and has been the state of the art in the field of speech recognition since then.

2.3 Formants and Linear Predictive Coding (LPC)

The vocal tract can be thought of as a vibrating cavity, or resonator, whose geometry (morphology) is constantly changing in continuous speech. This change in the geometry allows the superposition of different modes of vibration, and thereby different sounds.

Experiments have shown that for English vowels there are some characteristic frequencies of vibration called formants (Hillenbrand, Getty, Wheeler, & Clark, 1994; Hunter & Kebede, 2012; Deterding, 2006), that correspond to maxima of vibration, i.e. where the acoustic energy is focused. The first formant is roughly located from 0 to 1kHz, while the second formant is located from 1kHz to 2kHz, and so on (these frequency boundaries are not rigid, because there might be some second formants below 1kHz, as seen in the references). In the studies mentioned only the two first formants are taken into account for characterising each vowel. However, there are studies like (Prca & Ilić, 2010) in which 3 formants are used to characterise the Serbian vowels in order to improve accuracy when performing classification of vowel sounds.

Formants can be found by performing an acoustic power spectrum of a sound signal (or a spectrograph of the signal), and identifying the peaks in the spectrograph. This can be done by computing the envelope of the signal's frequency spectrum. Since the envelope contains information about the energy peaks, the formants can be determined. To predict the envelope of a signal $s[n(\Delta t)]$ sampled in discrete time steps $n = 0, 1, 2, \dots$, the Linear Predictive Coding method was introduced (Deng & O'Shaughnessy, 2003). The goal of LPC is to predict $s[n]$ with a linear combination of $s[n-1], s[n-2], \dots, s[n-M]$, where M is the integer that sets the number of predicting signal samplings.

2.4 Mel Frequency Cepstral Coefficients (MFCC)

This method aims to mimic the procedure performed by the human hearing system to decode a sound signal. It can be summarised in the steps shown in table 1 (Lyons, n.d.).

Table 1: Steps of the MFCC method compared to the sound analysis of the human body.

Step	Human Detection	Mel Frequency Coefficients
1.	The hearing sense is sensible to the tonal information change as a word or tune progresses.	Frame the signal into short frames to account that on short time scales the audio signal does not change much.
2.	Different frequencies are detected by vibrations at different spots of the cochlea depending on the frequency of the incoming sounds.	For each frame calculate the periodogram estimate of the power spectrum.
3.	The cochlea cannot discern the difference between two closely spaced frequencies.	Apply the Mel filterbank to the power spectra, sum the energy in each filter.
4.	Loudness is detected on a non-linear scale, where differences in small frequencies are more considerable.	Take the logarithm of all filterbank energies.
5.	—————	Take the DCT (Discrete Cosine Transform) of the log filterbank energies.
6.	Identifies the linguistic content and discards all the other stuff (background noise, emotion)	Keep DCT coefficients 2-13, discard the rest.

2.5 Decision Tree Classifiers (DTC)

Classifiers are computer tools that learn patterns from labelled data, so that when a new sample is presented to the computer, it will classify the sample into one of the different labels. In other words, consider a data set $D = \{(d_i, l_i)\}$, where d_i is a vector of N components (features). The vector d_i is also called a sample. l_i is a label. The classifier is a function $f(d) : F \rightarrow L$, where F is the vector space of the samples and L the space of the labels.

To see how DTCs work, consider the following example. Suppose that a data set consists of samples in one dimension: each sample is the salary of every single person in Colombia. The labels for this data set are the social stratum. We might have people that earn a lot of money, but they live in stratum 2, or vice versa. However, these cases are strange and in general there is a structure that allows predicting the label (social stratum) from the samples. The DTC will try to learn the salary ranges for each single social stratum, so that when a new sample is presented to the model, it will classify it in a specific label. The boundaries of the salary ranges are random at first, but as the learning process advances, the boundaries become more accurate and clear.

To test how well the DTC is able to predict labels, a data set D' called the testing data is normally built. D' consists of M samples with known labels. Only the samples, and not the labels, are presented to the DTC. The percentage of correctly predicted labels will tell the accuracy or effectiveness of the DTC. For a more thorough explanation, refer to (Tan, Steinbach & Kumar, 2005).

3 Methodology

The pipeline of this study (see Figure 1) was to extract and prepare audio files from the Vocal Joystick Vowel Corpus. Then, formants and cepstral coefficients were computed

for each file. The set of files with their corresponding formants and cepstral coefficients was split into training and testing data. A DTC was trained with the training data, and tested with the testing data, yielding an accuracy percentage. The characteristics of each step are explained in this section.

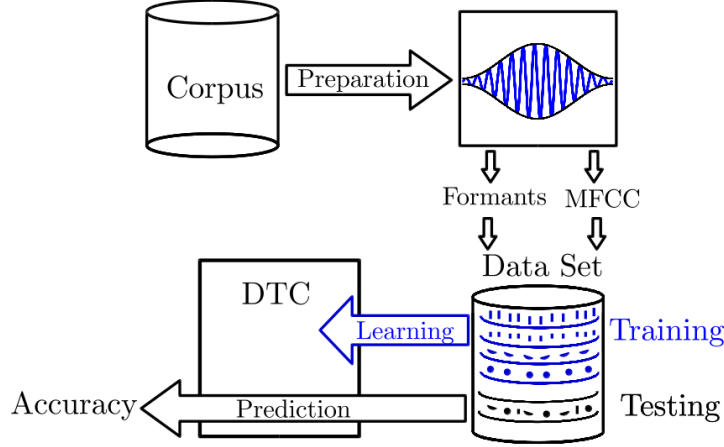


Figure 1: Diagram of the method used to measure the performance of formants and MFCC in pronunciation prediction.

3.1 Corpus and Audio Processing

In this study we used the Vocal Joystick Vowel Corpus (Bilmes, Wright, Xiao, Malkin Kilanski, 2006). This project selected a group of nine monophthongs and 12 vowel-to-vowel transitions. Sounds with different duration, amplitude and intonation were recorded for each vowel:

- Duration: short (1 second), long (2 seconds) and nudge (very short repetitions of the same vowel).
- Amplitude: quiet, normal, loud, quiet to loud and loud to quiet.
- Intonation: level, rising and falling.

From this corpus, we considered all the sound files available in the corpus corresponding to the phonemes /ə/, /æ/, /ɑ:/ and /ʌ/ whose duration was either short or long, whose amplitude was quite, normal or loud, and whose intonation was in a single level. A total of 784 files satisfied these conditions.

3.2 Data Preparation

An example of a raw data file from the corpus is shown in Figure 2a. It can be seen that there are silent parts and the amplitude is not normalised. Also, even though we selected a single level of intonation, it is clear that the amplitude of the sound decreases as time progresses. Therefore, for each file we deleted the silent part. Then, a section of duration 40ms was selected so that the signal does not change considerably and is approximately periodic (Figure 2b). Finally, a Hamming window was applied to the resulting sound signal with the purpose of smoothing the Fourier Transform used in the MFCC extraction (Figure 2c).

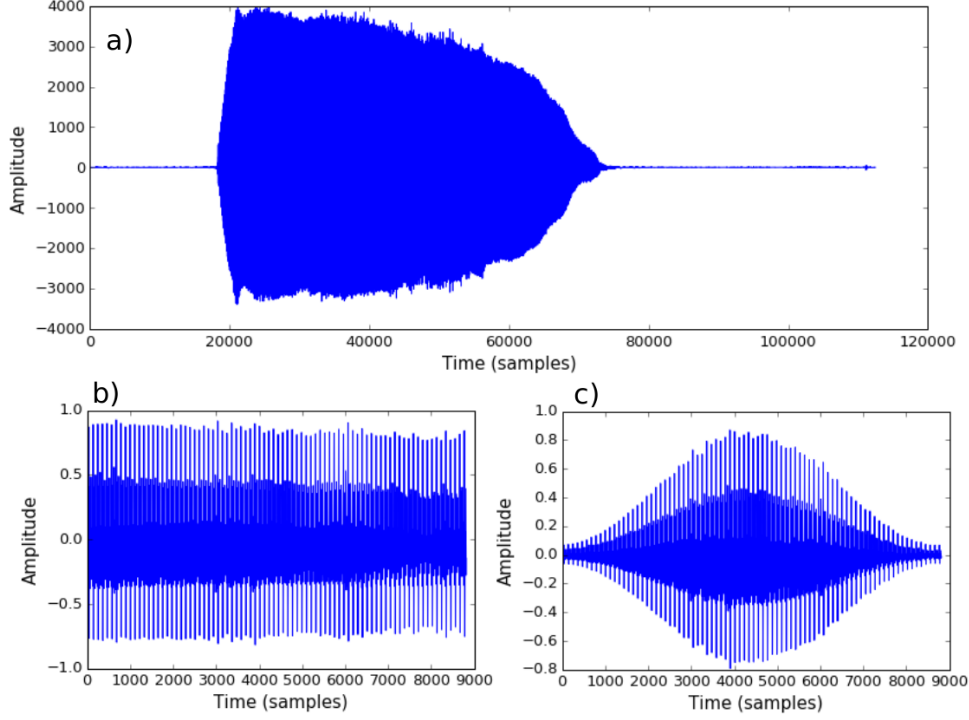


Figure 2: Preparation example of a sound signal corresponding to the $/\Lambda/$ phoneme. a) Raw signal. b) Silence deletion, amplitude normalisation and selection of a 40ms interval. c) Application of a Hamming window.

3.3 Evaluation of Formants and MFCC Performance with a DTC

Then, the LPC Python implementation by Danilo Bellini was used to compute the formants (Bellini, 2016). To do the calculation of the MFCC, the Python implementation by James Lyons (Lyons, 2016) was used. Finally, the set of computed formants and the set of computed cepstral coefficients were partitioned into two groups each. One of the groups contained two thirds of the data, and the other one contained a third. The former was the training data, and the later was the testing data. To exemplify this procedure, denote $S = \{(s, v)\}$ as the set of the sound signal files, each of which is labelled by a vowel v . Let $F(s)$ be a function that computes the formants F of a sound signal s . Let $X = \{(F(s), v) \forall s \in S\}$ be the data set of formants, labelled by their respective vowel. Let $X_{tr} \subset X$ be the set of training data and $X_{te} \subset X$ be the set of testing data. The DTC learns the ranges corresponding to each vowel in the formants space by creating a function $DTC(F(s))$ that takes the values of the labels, i.e. the phonemes $/\partial/$, $/\text{æ}/$, $/\alpha:/$ and $/\Lambda/$. For instance, if after the learning process is finished, a sample $F(s)$, labelled by the phoneme $/\partial/$, is presented to the DTC, and $DTC(F(s)) = /æ/$, then the DTC failed to predict the correct phoneme. If instead $DTC(F(s)) = /\partial/$ then the DTC was successful in the prediction. The accuracy is thus defined as the percentage of correctly predicted phonemes from the testing data. We used the accuracy to test the quality of prediction by the DTC using formants and Mel coefficients separately. This gives a measure of how well could formants and cepstral coefficients help in the objective evaluation of pronunciation.

4 Results, Analysis and Discussion

4.1 Formants

From the 784 sound files, 142 were discarded for this analysis and for the one corresponding to the MFCC. The reason to discard these files was that some recorded sounds contained parts in which the signal was not periodic due to a trembling voice from the speaker. This could cause the LPC and MFCC algorithms to be affected. The criterion used to discard the data is now explained. Denote the recordings or sound signals that belong to vowel v as (s, v) . Also, denote the mean value of the i -th formant ($i = 1, 2$) and the standard deviation of the i -th formant, for a vowel v , as $(\mu_i, v), (\sigma_i, v)$, respectively. The remaining 642 sound signals satisfied the condition

$$|F_i(s) - \mu_i| < 1.5\sigma_i, \quad (1)$$

for each vowel v , where F_i refers to the i -th formant. The computed formants for these sound signals are shown in figure 3.

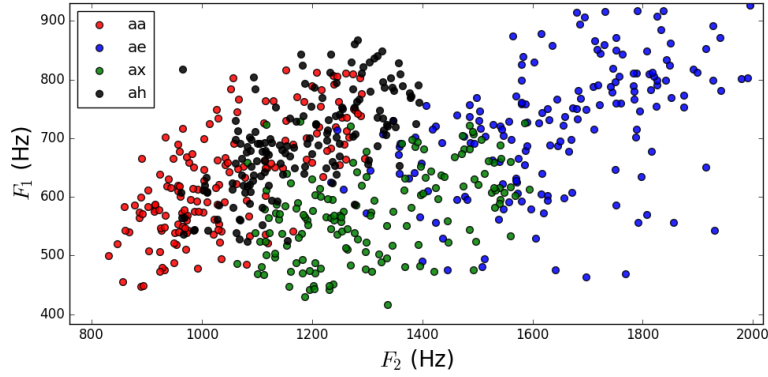


Figure 3: Frequencies in Hertz of the first two formants for the studied phonemes. aa is /a:/, ax is /ə/, ae is /æ/ and ah is /ʌ/.

The DTC trained with the formants yielded a 70% accuracy. Although this success rate of the predictive model is a very low rate, it can be considered as good taking into account that only two features were used to predict the data: the first two formants. This means that with little information about four very similar sounding phonemes, we were able to predict, with an accuracy of 70%, the vowel of a sound taken at random from the recordings obtained from the corpus.

Furthermore, Figure 3 shows a non-Gaussian distribution in either formant for every phoneme. This is supported by observations in which it was seen that formants do not only depend on the geometry of the vocal tract of each person, but also are statistically dependent on the age, as shown in (Hawkings & Midgley, 2005); as well as on gender, as shown in (Hillenbrand, Getty, Clark, & Wheeler, 1995). The formants measured are shown in table 2 and compared with measurements from other studies.

Table 2: Formants measurement in Hertz. Our measurements are compared with Hawkings *et. al.*, Hunter *et. al.*, Deterding.

Study	Our		Hawkings <i>et. al.</i>		Hunter <i>et. al.</i>		Deterding	
Phoneme	F_1	F_2	F_1	F_2	F_1	F_2	F_1	F_2
/ə/	631	1049	496	833	643	1019	625	973
/æ/	720	1644	696	1574	667	1565	748	1360
/ɑ:/	573	1311	608	1062	680	1193	757	1211
/ʌ/	693	1182	629	1160	661	1296	724	1282

Studies shown in Table 2 show that formants vary a lot with age, gender and geography. Therefore, we see that identifying a phoneme by its formants is not trivial, but as we showed, can be done with an estimated accuracy of 70%. Despite the low accuracy, formants keep being used for vowel pronunciation research, as in (Rasilo & Räsänen, 2017), in which the process of infants learning their native language was simulated through a Learning Virtual Infant that received sounds from its caregivers. At first the Learning Virtual Infant babbled randomly, but as the learning process advanced, it began to compute formants to identify similar sounding vowels and to improve its pronunciation.

4.2 MFCC

13 MFCC were obtained for every sound signal. In order to visualise the MFCC of every sound signal, a dimension reduction from 13 to 2 (i.e. from the MFCC space to a plane) was performed with Principal Component Analysis (PCA). This method projects the data from the high-dimensional space onto a plane that preserves the maximum possible variance of the data. The distribution of the MFCC for the different phonemes on the plane computed with PCA is shown in Figure 4.

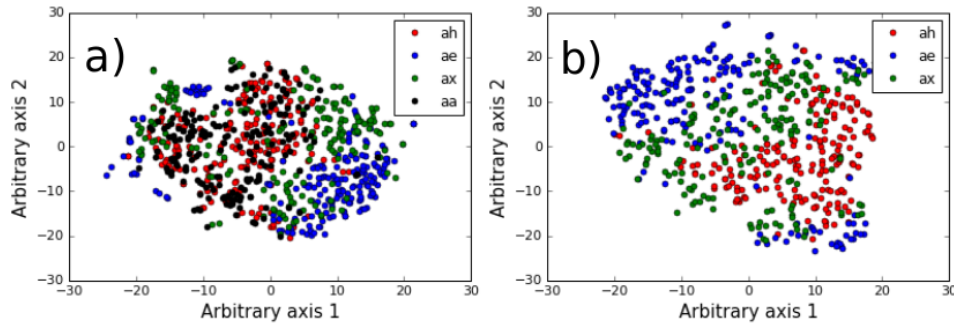


Figure 4: PCA of the MFCC. a) shows the four studied phonemes and b) shows only /a:/, /æ/ and /ʌ/.

As it can be seen from Figure 4a, the MFCC of the phonemes seem mixed. In particular, note that the points corresponding to the phoneme /a:/ are mixed with the ones corresponding to the phoneme /ʌ/. This can also be seen in the formants figure. Ignoring the /a:/ points produces Figure 4b, in which the clusters of data can be identified.

Partitioning the data shown in Figure 4a in training and testing data, an accuracy of 56% was accomplished with the DTC. While if the data shown in Figure 4b was used, an accuracy of 71% was measured. These very low accuracies show that the MFCC method lacks sensibility to subtle differences between the studied sound signals.

But, why are the MFCC so good at speech recognition then? The first thing to mention is that the MFCC are normally used to train models that are more complex than a DTC, like neural networks. The goal in speech recognition is to understand words, instead of single sounds. To check the accuracy of the MFCC at recognising different sounding sounds, we proceeded to analyse the phonemes /ʌ/, /e/, /u/ and /o/. The PCA results are shown in Figure 5.

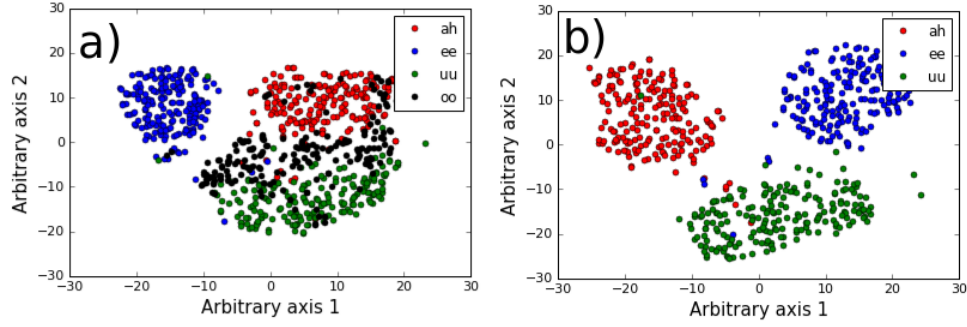


Figure 5: PCA of the MFCC. a) shows the phonemes $/\Lambda/$, $/e/$, $/u/$ and $/o/$. b) shows only $/\Lambda/$, $/e/$ and $/u/$. ee is $/e/$, uu is $/u/$ and oo is $/o/$.

Using the data shown in Figure 5a, the accuracy was measured in 73%. Clearly the MFCC corresponding to the phoneme $/o/$ were mixed with the ones corresponding to the phoneme $/u/$. To avoid this noise, if $/o/$ is removed, an accuracy of 88% was reached. This shows that MFCC are a reasonable method to recognise different phonemes.

5 Conclusions and Perspectives

The Vocal Joystick Corpus was used to build a data set of sounds corresponding to the phonemes $/ə/$, $/æ/$, $/ɑ:/$ and $/\Lambda/$. From each of the sound signals, both formants and MFCC were computed. These were used to train a DTC that acted as a classifier. Using formants, an accuracy of 70% was achieved. Using MFCC, an accuracy of 52% was measured. MFCC's performance was much better when the DTC was trained with the coefficients computed from the phonemes $/\Lambda/$, $/e/$ and $/u/$, reaching 88%. However, these speech recognition tools fail to correctly predict the phonemes of similar sounding signals.

As a response to this deficiency, computer and linguistics scientists work to build computational tools that learn subtle differences between sounds in order to assess a person's pronunciation correctly. For instance, a research in the Chinese language showed results that improved significantly the previous recognition models in detecting mispronunciation of phonemes (Wei, Hu, Hu, & Wang, 2009). However, the advance of technology to support pronunciation learning must be accompanied by designed learning programs that help students to learn autonomously. More people need to be encouraged to work in this interdisciplinary field that is pioneering and improving language teaching.

References

- [1] Baker, J., Li Deng, Glass, J., Khudanpur, S., Chin-hui Lee, Morgan, N., & O'Shaughnessy, D. (2009). Developments and directions in speech recognition and understanding, Part 1 [DSP Education]. IEEE Signal Processing Magazine, 26(3), 75-80. <http://dx.doi.org/10.1109/msp.2009.932166>
- [2] Bellini, D. (n.d.). Expressive Digital Signal Processing (DSP) package for Python, 2016. Retrieved November 16, 2016, from <https://github.com/danilobellini/audiolazy>
- [3] Bilmes, J., Wright, R., Xiao, L., Malkin, J. & Kilanski, K. 2006. <http://melodi.ee.washington.edu/vj/index.html>

- [4] Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions On Acoustics, Speech, And Signal Processing*, 28(4), 357-366. <http://dx.doi.org/10.1109/tassp.1980.1163420>
- [5] Deng, L. & O'Shaughnessy, D. (2003). *Speech processing* (1st ed., pp. 19-26). New York: Marcel Dekker.
- [6] Deterding, D. (2006). The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *Journal Of The International Phonetic Association*, 36(02), 187. <http://dx.doi.org/10.1017/s0025100306002544>
- [7] Ghaffarvand Mokari, P. & Werner, S. (2016). Perceptual assimilation predicts acquisition of foreign language sounds: The case of Azerbaijani learners' production and perception of Standard Southern British English vowels. *Lingua*, 185, 81-95. <http://dx.doi.org/10.1016/j.lingua.2016.07.008>
- [8] Hawkins, S. & Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal Of The International Phonetic Association*, 35(02), 183. <http://dx.doi.org/10.1017/s0025100305002124>
- [9] Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal Of The Acoustical Society Of America*, 97(5), 3099. <http://dx.doi.org/10.1121/1.411872>
- [10] Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *Recall*, 15(01). <http://dx.doi.org/10.1017/s0958344003000211>
- [11] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., & Jaitly, N. et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. Retrieved from <http://static.googleusercontent.com/media/research.google.com/en/pubs/archive/38131.pdf>
- [12] Hunter, G. & Kebede, H. (2012). Formant frequencies of British English vowels produced by native speakers of Farsi. *Société Française d'Acoustique*. Retrieved from <https://hal.archives-ouvertes.fr/hal-00810580/document>
- [13] Lee, K. (1988). On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, 7(4), 375-379. [http://dx.doi.org/10.1016/0167-6393\(88\)90053-2](http://dx.doi.org/10.1016/0167-6393(88)90053-2)
- [14] Lyons, J. Mel Frequency Cepstral Coefficient (MFCC) tutorial, Practical Cryptography. Retrieved October 15, 2016, from <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfcc/>
- [15] Martin, I. & Jackson, C. (2016). Pronunciation Training Facilitates the Learning and Retention of L2 Grammatical Structures. *Foreign Language Annals*, 49(4), 658-676. <http://dx.doi.org/10.1111/flan.12224>
- [16] Mirzaei, K., Gowhary, H., Azizifar, A., & Esmaeili, Z. (2015). Comparing the Phonological Performance of Kurdish and Persian EFL Learners in Pronunciation of English Vowels. *Procedia - Social And Behavioral Sciences*, 199, 387-393. <http://dx.doi.org/10.1016/j.sbspro.2015.07.523>

- [17] Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5), 393-408. <http://dx.doi.org/10.1080/09588220802447651>
- [18] Oppenheim, J. & Magnasco, M. (2013). Human Time-Frequency Acuity Beats the Fourier Uncertainty Principle. *Physical Review Letters*, 110(4). <http://dx.doi.org/10.1103/physrevlett.110.044301>
- [19] O'Grady, W. & Dobrovolsky, M. (1997). *Contemporary Linguistics* (3rd ed., pp. 40-87). Boston: Bedford/St. Martin's.
- [20] Prica, B. & Ilić, S. (2010). Recognition of vowels in continuous speech by using formants. *Facta Universitatis - Series: Electronics And Energetics*, 23(3), 379-393. <http://dx.doi.org/10.2298/fuee1003379p>
- [21] Rasilo, H. & Räsänen, O. (2017). An online model for vowel imitation learning. *Speech Communication*, 86, 1-23. <http://dx.doi.org/10.1016/j.specom.2016.10.010>
- [22] Rosen, S. & Howell, P. (2011). *Signals and systems for speech and hearing* (2nd ed., p. 163). London: Academic Press.
- [23] Sethna, J. (2006). *Statistical mechanics: Entropy, Order Parameters and Complexity* (1st ed., p. 299). Oxford: Oxford University Press.
- [24] Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed., p. 150). Boston: Pearson.
- [25] Wei, S., Hu, G., Hu, Y., & Wang, R. (2009). A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models. *Speech Communication*, 51(10), 896-905. <http://dx.doi.org/10.1016/j.specom.2009.03.004>
- [26] Wellekens, C. (1998). *Introduction to Speech Recognition Using Neural Networks*. Bruges: ESANN. Retrieved from <https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es1998-456.pdf>
- [27] Zegers, P. (1998). *Speech Recognition Using Neural Networks* (Master of Science with a major in Electrical Engineering). University of Arizona.